S&DS 363 Final Project

April 2020

Introduction

The World Happiness Report is a leading survey in assessing the state of global happiness. There has been a report ever since the inception of the survey in 2012. In recent years, the report has gained global recognition as governments, organizations and civil societies increasingly use happiness indicators to inform their policy-making decisions. Leading experts across various fields such as economics, psychology, survey analysis, national statistics, health and public policy describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

Design and Primary Questions

As there are now more than 7 reports that have been released, we chose to focus our analysis on the one published in 2015, as it included the most variables than any other report.

Given that dataset we will try to answer three different questions in this analysis using three different multivariate statistical techniques.

Firstly, along what lines do countries tend to be different when we examine factors that may contribute to their happiness score? With the help of Principal Component Analysis (PCA) we will be able to reduce all of the variables found in the model to a lower dimensional space that will allow us to make concrete comparisons.

Secondly, which countries are like other countries when it comes to considering the factors that affect their happiness scores? We believe that this grouping of countries will be best done through Cluster Analysis.

And finally, are there any latent variables that can best explain the differences between happiness scores among countries? In order to achieve this we will be using Factor Analysis which can tell us how many latent factors can explain the data and from there we will be able to deduce what exactly these factors are.

Data

The content of the report consists of happiness scores for each country along with all the explanatory variables that may affect the value of that score. The scores of the report use data from the Gallup World Poll and are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2014 and use the Gallup weights to make the estimates representative.

The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

Dystopia is an imaginary country that has the world's unhappiest people. The purpose in establishing Dystopia is to have a benchmark against which all countries can be favorably compared (no country performs more poorly than Dystopia) in terms of each of the six key variables, thus allowing each sub-bar to be of positive width. The lowest scores observed for the six key variables, therefore, characterize Dystopia.

More specifically for each of the variables:

-**Country**: Name of the country. Factor Variable.

-**Region**: Region that a country belongs to. Factor Variable.

-**Happiness Rank**: Rank of the country based on the Happiness Score. Factor Variable.

-**Happiness Score**: A metric measured in 2015 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest." Continuous Variable.

-**Standard Error**: The standard error of the happiness score. Continuous Variable.

-**Economy (GDP per Capita)**: The extent to which GDP contributes to the calculation of the Happiness Score. Continuous Variable.

-**Family**: The extent to which Family contributes to the calculation of the Happiness Score. Continuous Variable.

-**Health (Life Expectancy)**: The extent to which Life expectancy contributed to the calculation of the Happiness Score. Continuous Variable.

-**Freedom**: The extent to which Freedom contributed to the calculation of the Happiness Score. Continuous Variable.

-**Trust (Government Corruption)**: The extent to which Perception of Corruption contributes to Happiness Score. Continuous Variable.

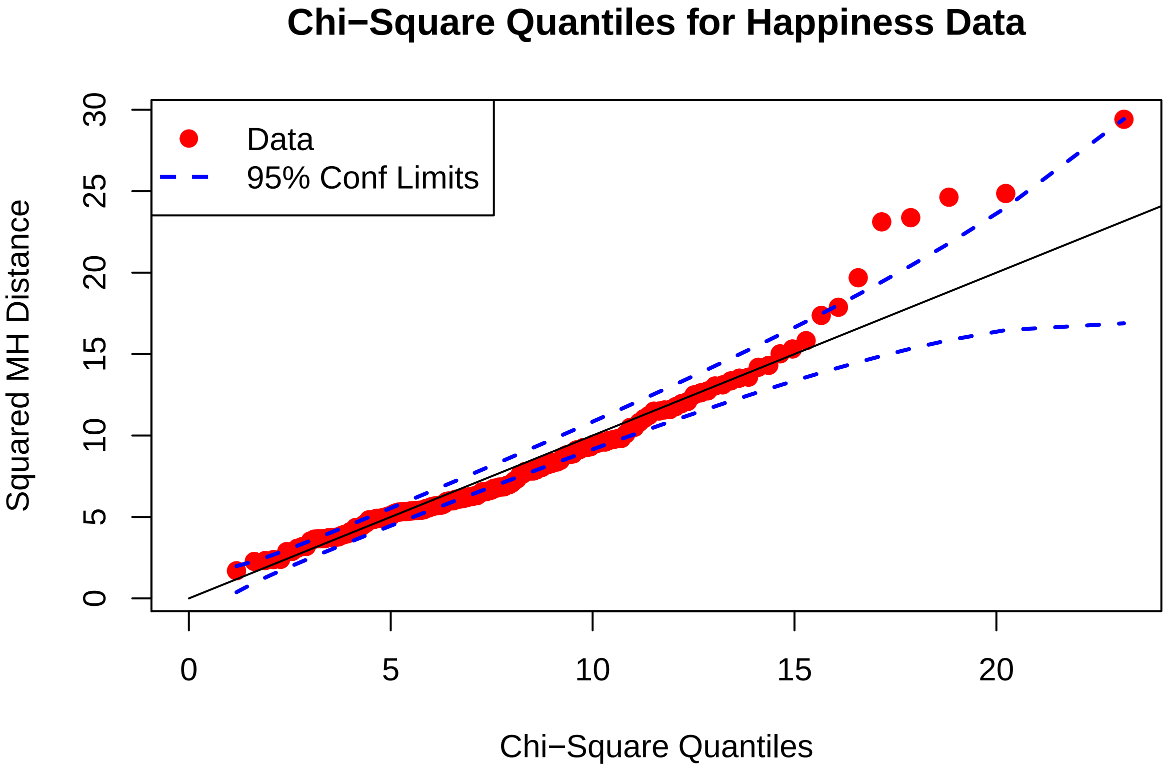
-**Generosity**: The extent to which Generosity contributed to the calculation of the Happiness Score. Contunuous Variable.

-**Dystopia Residual**: The extent to which Dystopia Residual contributed to the calculation of the Happiness Score. The residuals, or unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average 2014-2015 life evaluations. These residuals have an average value of approximately zero over the whole set of countries. Continuous Variable.

By adding all these factors listed above we get the happiness score for each of the countries.

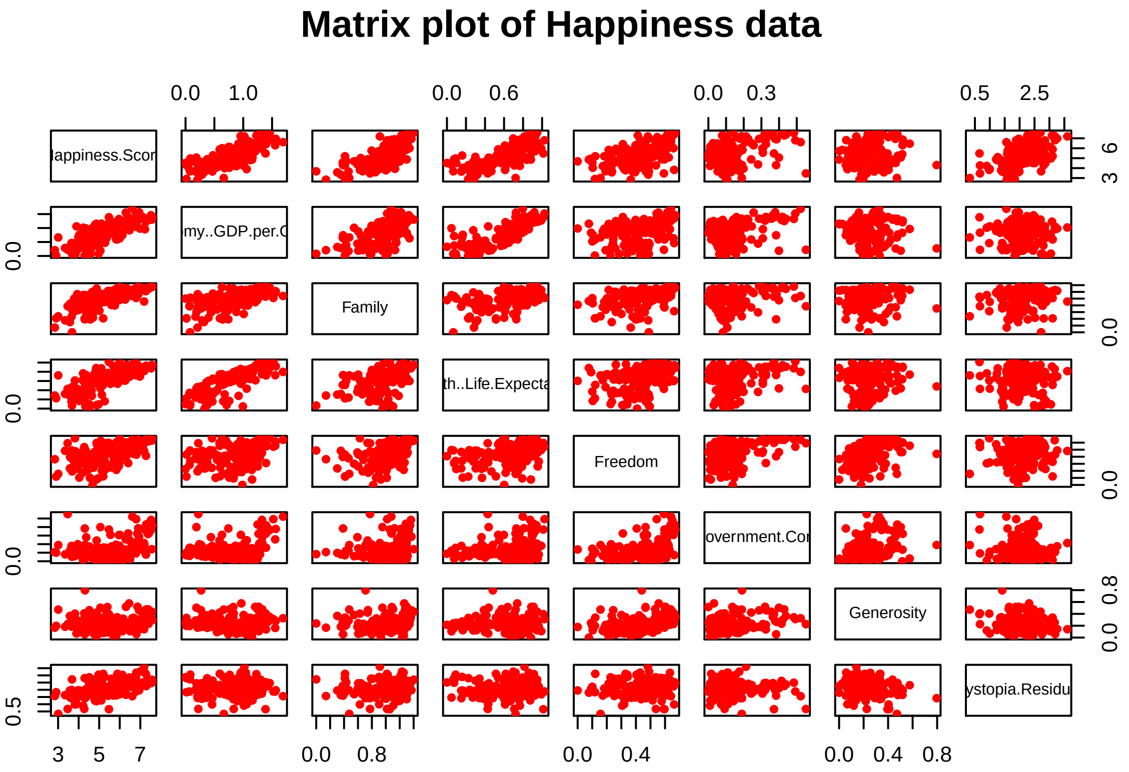
Multivariate Analysis

Principal Components Analysis



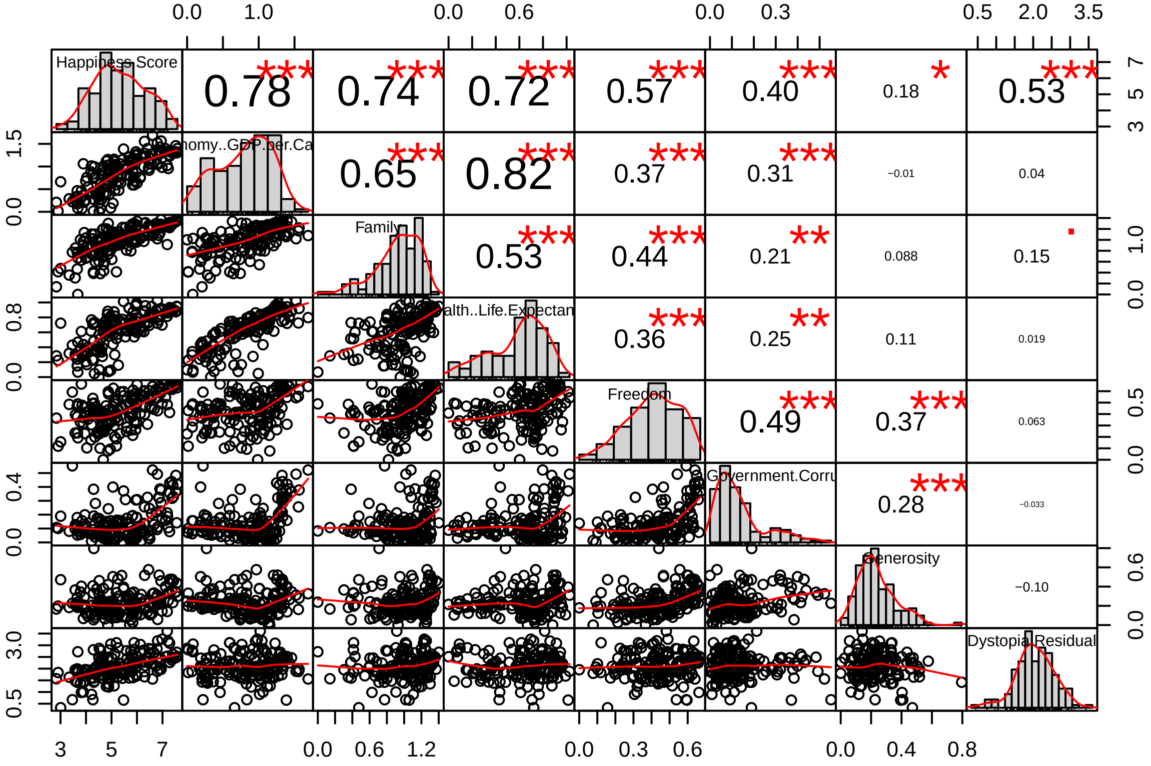
Looking at our chi-square quantile plot, we observe that our data is not perfectly normal as some of the data points are outside of the 95% confidence limits. Despite this, we feel that our data is still good enough for us to perform a PCA analysis on it.

```
plot(newdata, pch=19, cex=.7, col='red', main="Matrix plot of Happiness data")
```

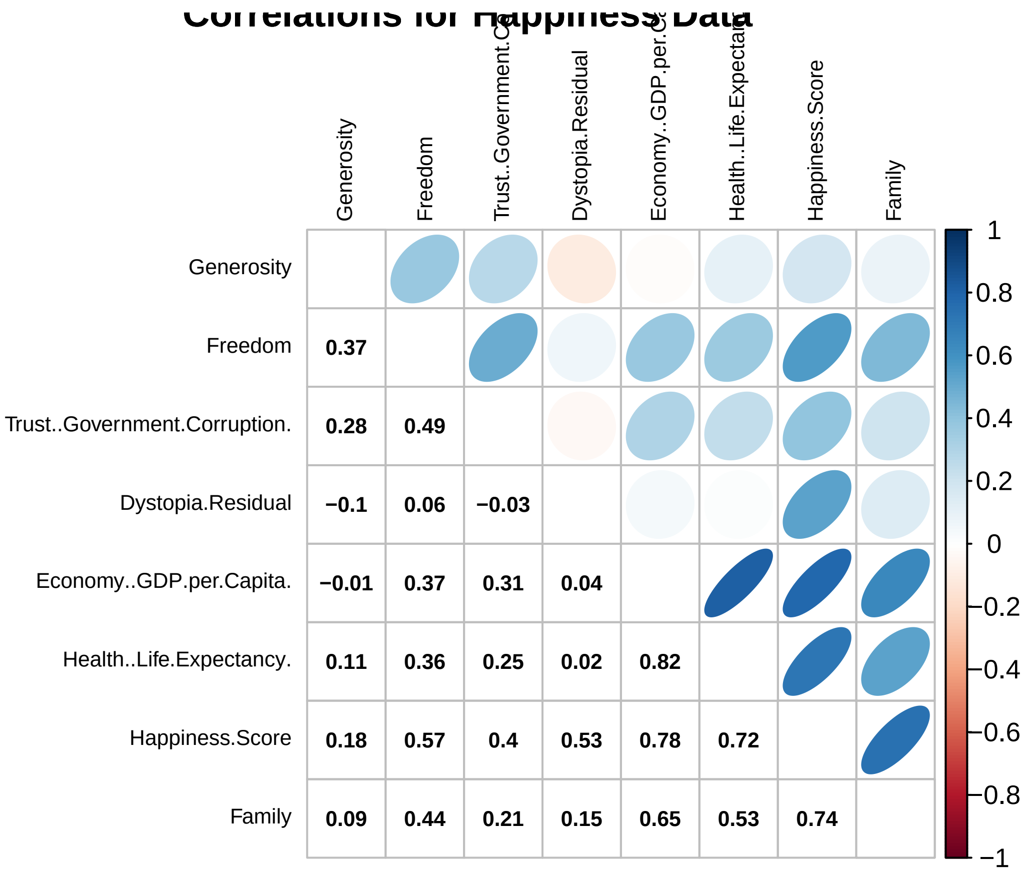


Matrix plot of Happiness data

```
chart.Correlation(newdata, histogram=TRUE, pch=19)
```



While checking for linearity in our data, we make a series of bivariate scatterplots to check for any non-linear relationships. Looking at our scatterplots, we can see that all of our variables have a linearly correlation. Because of the lack of any non-linear relationships, we can proceed with our PCA analysis.



Lastly, we decide to look into the correlation between our variables by creating a correlation table. From this, we determine that most of our variables are either strongly or moderately correlated with each other. Ideally, for a PCA analysis, we would want most of our variables to be moderately correlated with each other and our findings support this. Therefore, we can proceed with our PCA analysis.

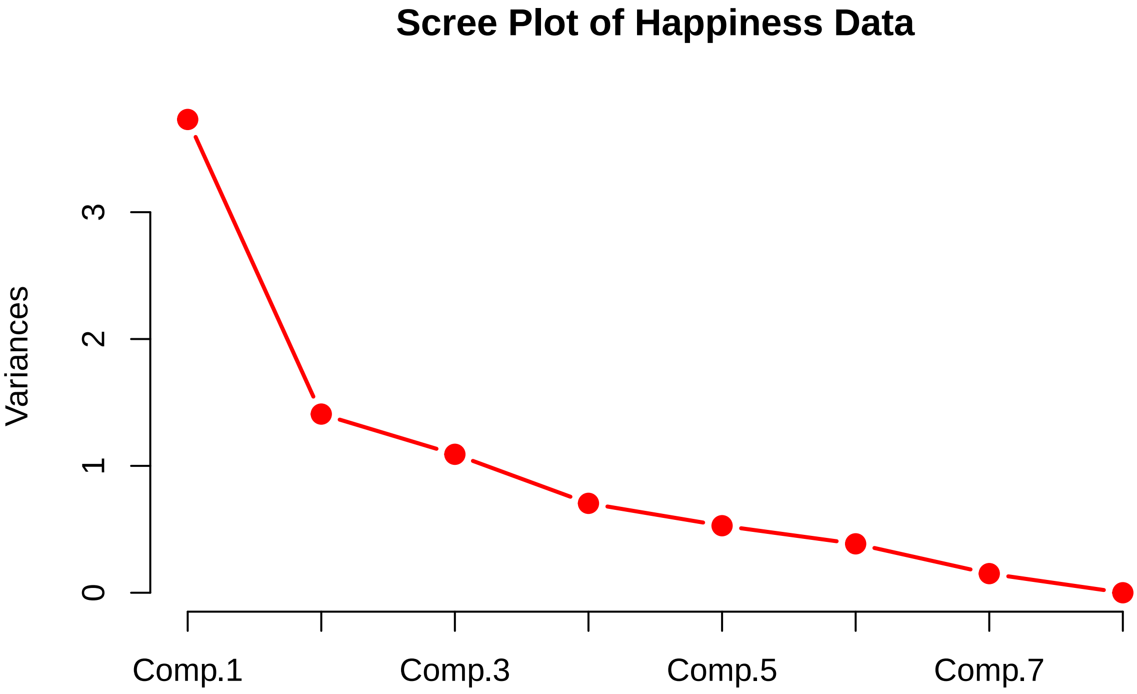
```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.9315070 1.1867517 1.0446560 0.83943558 0.72696380
## Proportion of Variance 0.4663399 0.1760474 0.1364133 0.08808151 0.06605955
## Cumulative Proportion  0.4663399 0.6423873 0.7788006 0.86688212 0.93294166
##                            Comp.6     Comp.7       Comp.8
## Standard deviation     0.62100735 0.38835102 1.965340e-04
## Proportion of Variance 0.04820627 0.01885206 4.828202e-09
## Cumulative Proportion  0.98114793 1.00000000 1.000000e+00
```

```
## Warning in if (loadings) {: the condition has length > 1 and only the first
## element will be used
```

```
## 
## Loadings:
##                               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Happiness.Score                0.49   0.17   0.22   0.04   0.09   0.02   0.07
## Economy..GDP.per.Capita.       0.44   0.18  -0.35  -0.05   0.17  -0.01   0.73
## Family                         0.41   0.15  -0.08   0.21  -0.59   0.57  -0.23
## Health..Life.Expectancy.       0.42   0.12  -0.35   0.13   0.42  -0.30  -0.61
## Freedom                        0.35  -0.37   0.19  -0.08  -0.51  -0.65   0.03
## Trust..Government.Corruption.  0.26  -0.45   0.10  -0.72   0.22   0.35  -0.12
## Generosity                     0.13  -0.61   0.22   0.64   0.30   0.19   0.12
## Dystopia.Residual              0.13   0.43   0.78  -0.02   0.20  -0.03  -0.02
##                               Comp.8
## Happiness.Score                 0.82
## Economy..GDP.per.Capita.       -0.29
## Family                         -0.19
## Health..Life.Expectancy.       -0.18
## Freedom                        -0.11
## Trust..Government.Corruption.  -0.09
## Generosity                     -0.09
## Dystopia.Residual              -0.39
```

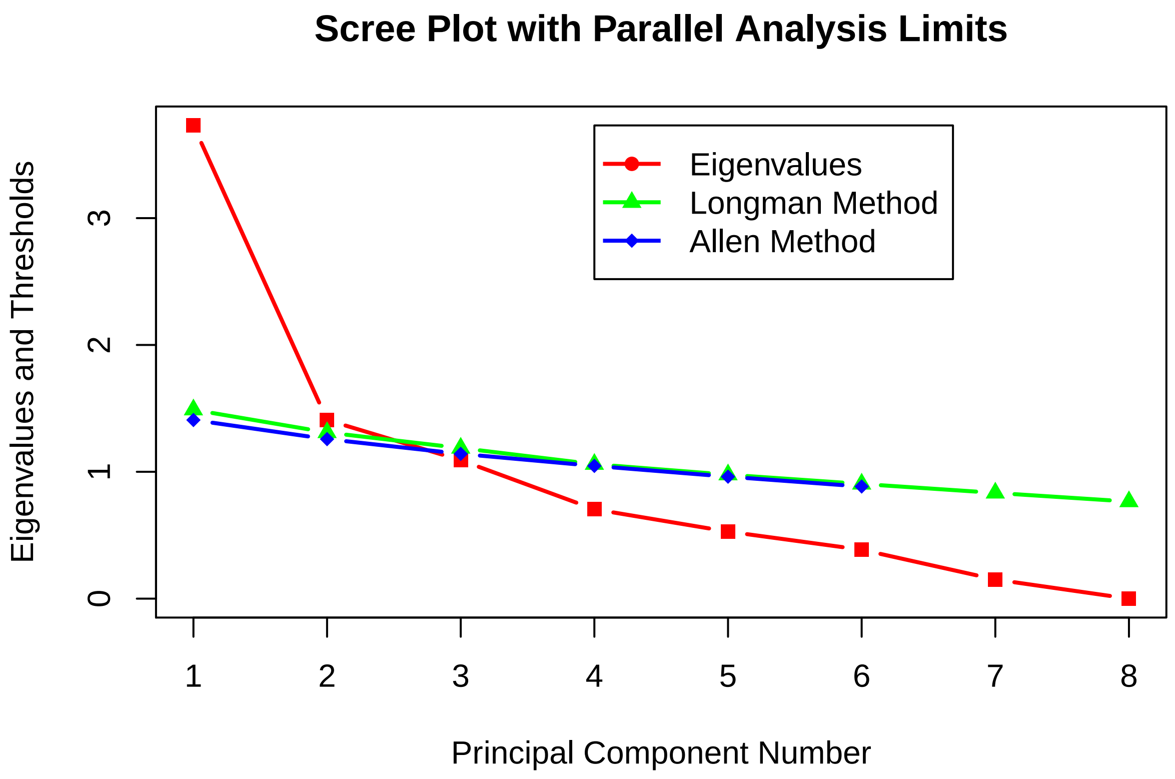
```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
##   3.73   1.41   1.09   0.70   0.53   0.39   0.15   0.00
```

Because we want to explain 80% of the variability in our data, we observe that we should keep 4 components, which explains about 86.6% of the total variance. If we use an eigenvalue > 1 rule, we see that we should keep 4 components since the fourth component has an eigenvalue that is less than 1.



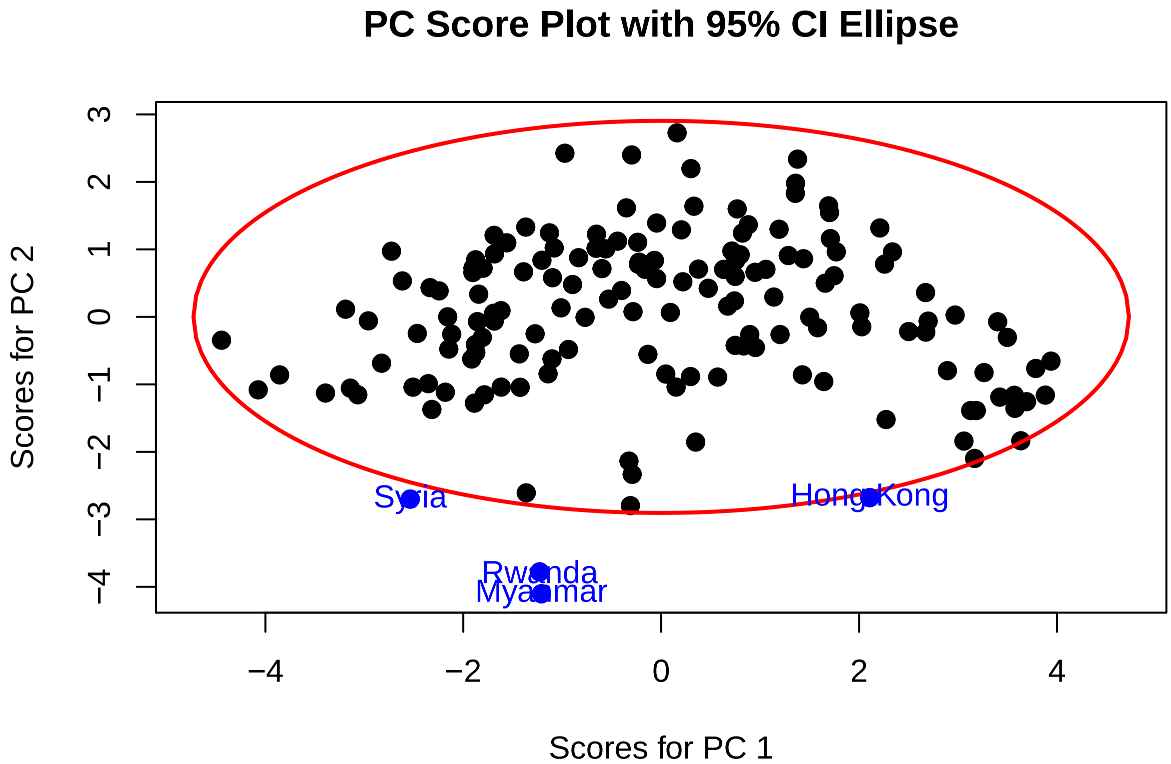
The scree plot indeed confirms our belief from earlier that the first 2-3 components explains most of the variability in our data. Between the 2nd and 3rd component, we see an elbow in our scree plot, which indicates that we should keep 2 components.

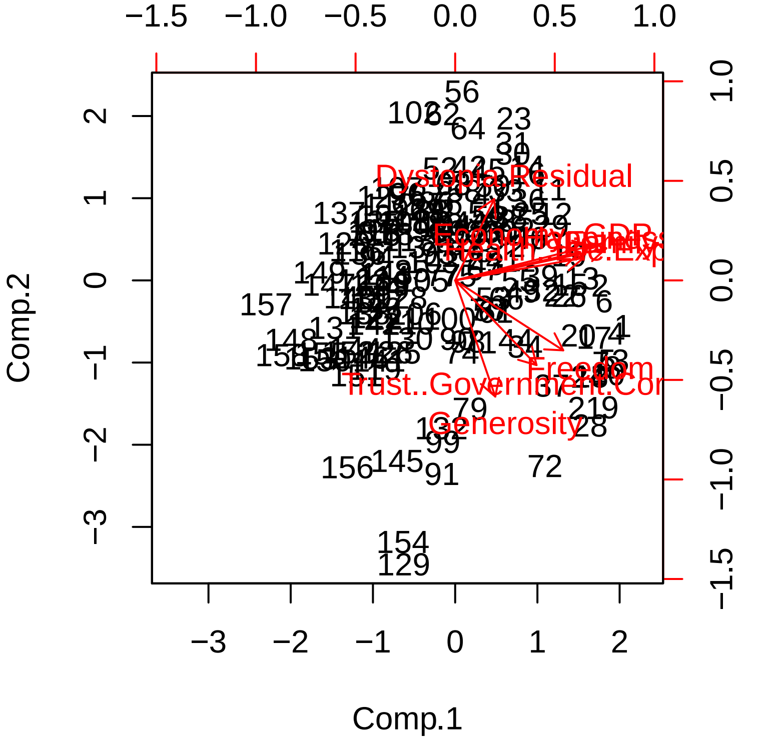
```
##   pcompnum  longman     allen
## 1        1 1.4891588 1.4083074
## 2        2 1.3099980 1.2575363
## 3        3 1.1872165 1.1413114
## 4        4 1.0597489 1.0465549
## 5        5 0.9771499 0.9615184
## 6        6 0.9040986 0.8840679
## 7        7 0.8335438        NA
## 8        8 0.7653181        NA
```



Finally, we decide to conduct a parallel analysis using the Longman and Allen method to determine the number of components we should keep. Looking at our parallel plot, we observe that only our first 2 eigenvalues are above the parallel threshold. Overall, we decide to keep 2 principal components from our analysis because the scree plot and parallel analysis are generally considered to be more reliable and desirable in determining the number of components to keep.

Our loading coefficients table shows that component 1 has a moderate correlation for all of the variables involved. We can then say that component 1 likely represents overall happiness levels in a country because of its consistent correlation with all of the variables. Looking at component 2, we see a strong negative correlation value for the variables ‘Generosity’, ‘Freedom’ and ‘Trust in Government’. We can then say that component 2 is picking up on variation relating to democracy and government trust.





Looking at our PC score plot, we see that most of the data points are within the 95% confidence interval ellipse. However, we do have a few outliers, particularly in the direction of the 2nd component. Recall that the 2nd component is related to democracy and government trust. Interestingly enough, we observe that 3 out of 4 of these countries outside the ellipse have, in the past decades, experienced some form of violence or genocide that largely impacted the country. This could explain why the happiness score for 'Syria', 'Rwanda', and 'Myanmar' is largely dependent on democracy and government trust.

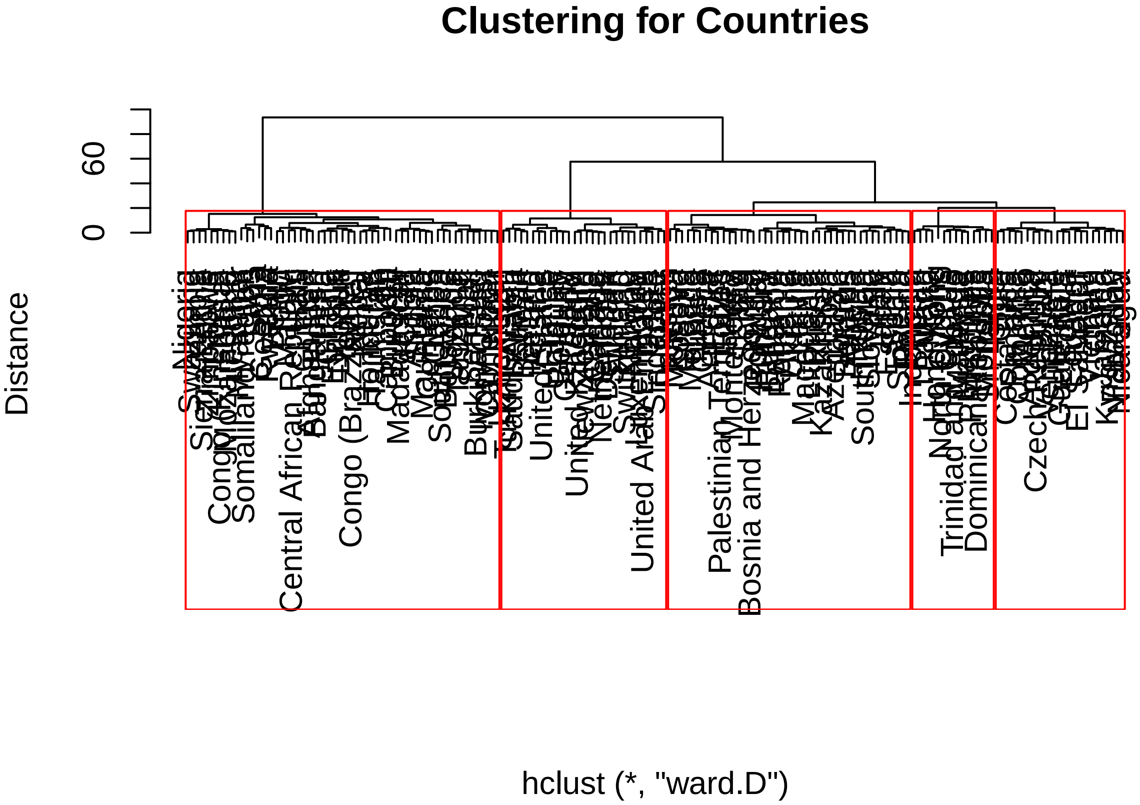
From the biplot, we can observe that all of the variables are pointing in the positive direction for component 1, which means that all of them are positively affecting happiness. However, for component 2, we see a divergence between variables related to life expectancy/wealth and variables relating to generosity and government trust. This divergence means that countries with higher GDP per capita and life expectancy depend less on generosity and government trust when it comes to their happiness levels. The same can be said vice versa for countries that largely depend on generosity and government trust for their happiness.

Cluster Analysis

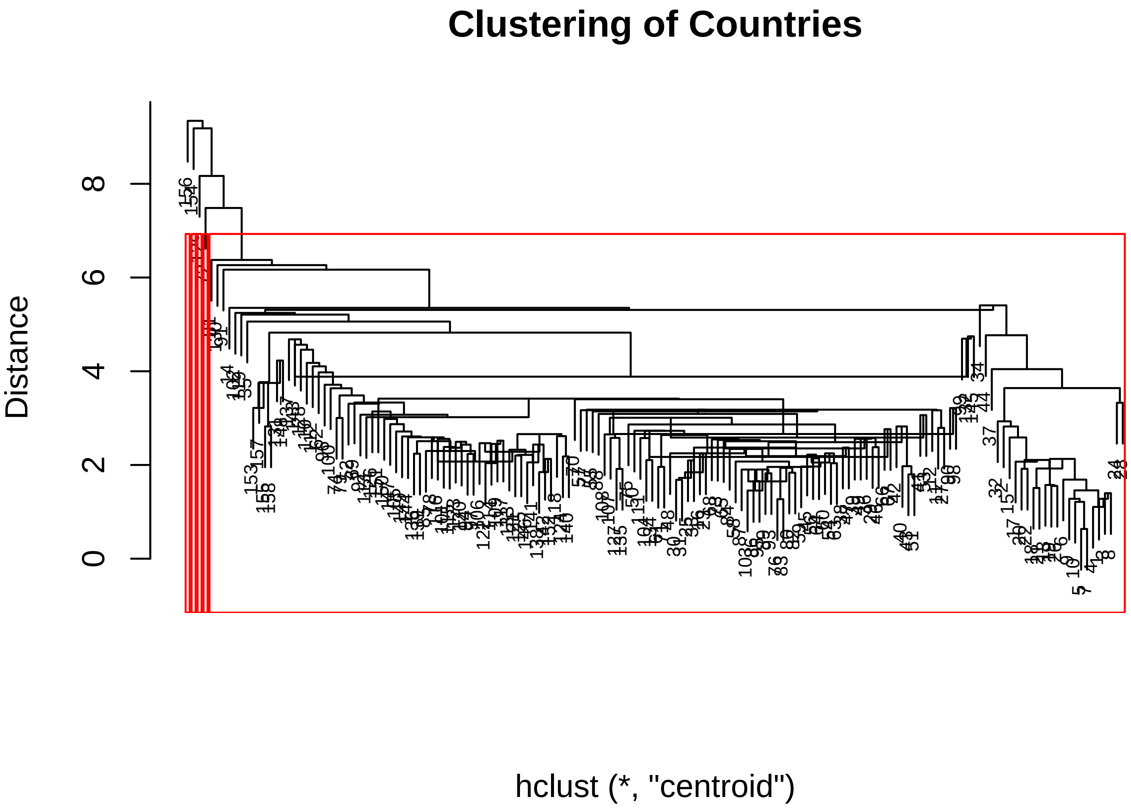
Since our data is continuous, we decided to do a cluster analysis using two methods to group the data. Our first method uses the euclidean distance and ward.D as our agglomeration method. In our second method, we instead use the manhattan distance metric and centroid as our agglomeration method.

```
newdata <-  scale(na.omit(newdata))
```

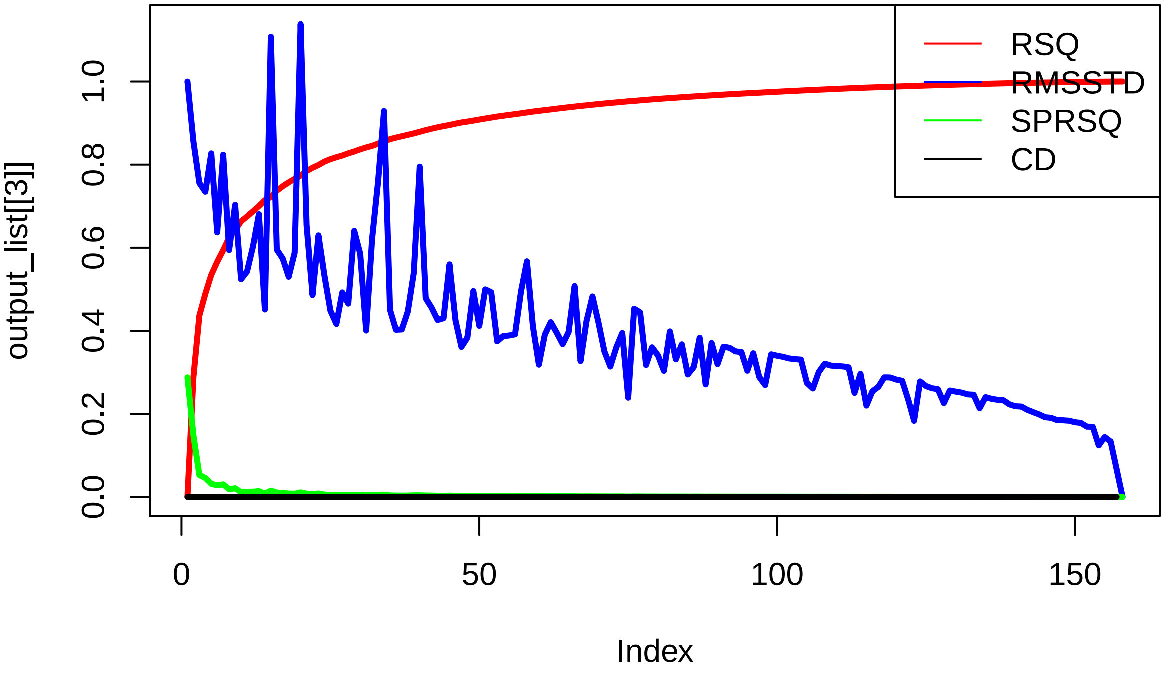
To start off, we decide to scale our data to make sure that no variable is scale dependent, meaning that some variables could have more influence over others due to their units.



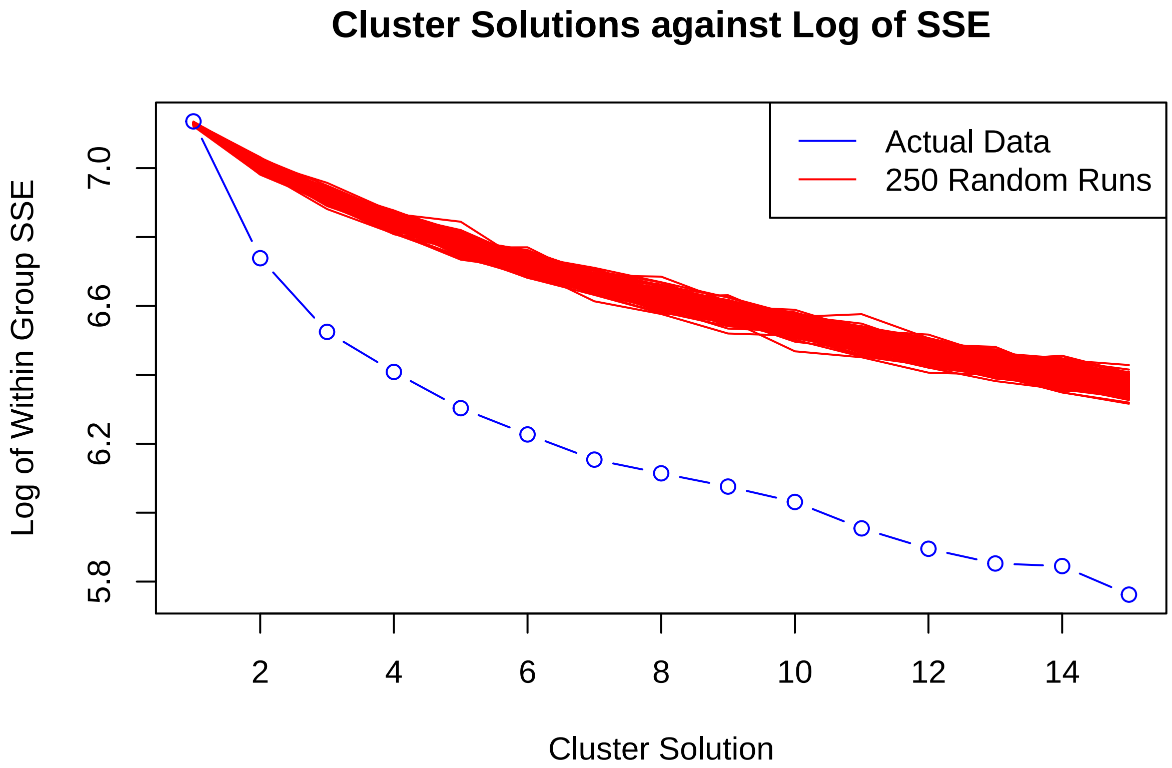
Next, we decide to cluster in 5 groups using the euclidean distance and ward.D agglomeration method. We can see that, although it is difficult to identify each individual country, our dendrogram tree is broken down pretty evenly in 5 separate clusters of countries.

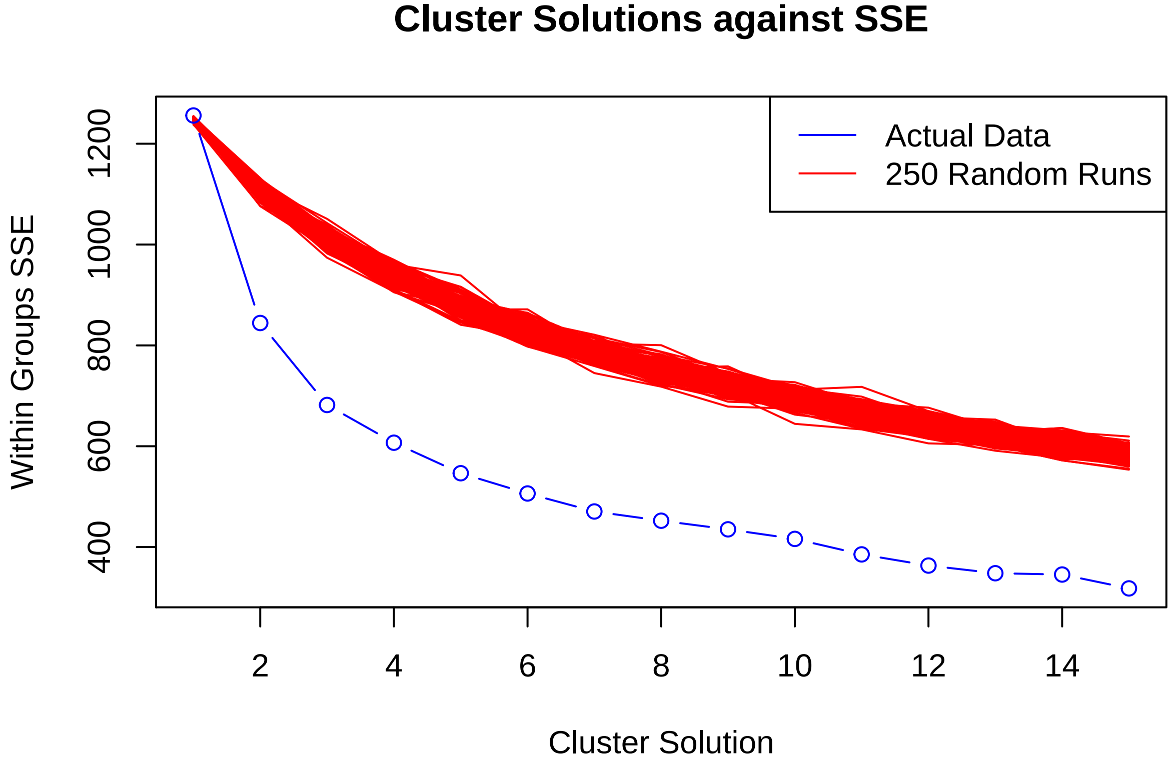


On the other hand, when using the manhattan distance and centroid agglomeration method, we see very different results. We encounter an uneven clustering of countries where a large majority of countries are clustered into one group leaving the other 4 cluster groups with only 1 or 2 countries. In addition, our dendrogram tree branches are all broken down into many separate sub-branches in a uneven and complex way.

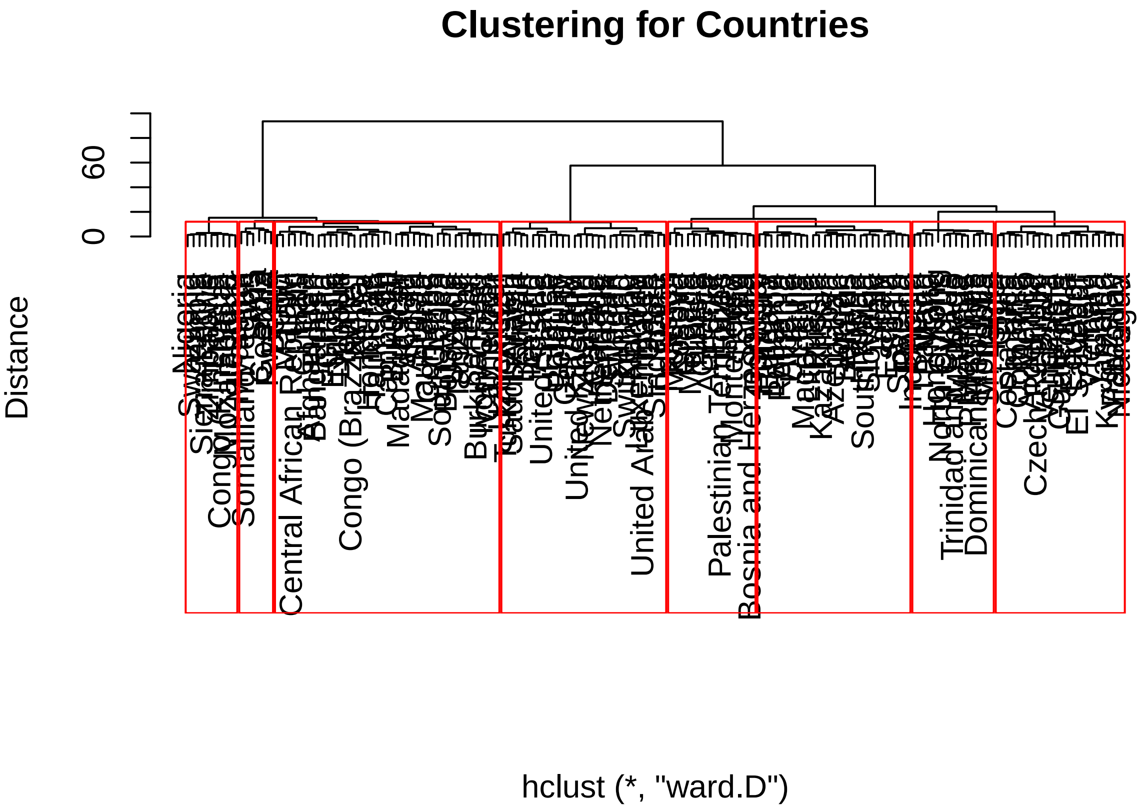


In the above plot, we use the euclidean distance and ward.D agglomeration method to plot the Rˆ2, which is a measure of how much variability in the data our model explains. We see that Rˆ2 slows to significantly slow down in its increase after about 8-10 cluster groups. The Semi-Partial R Squared, which measures the relative change in within-clusters Sum of Squares, seems to as well become near zero after 8-9 groups with significantly smaller marginal gains by the addition of one more group after that point.





In the graph above, which represents the log of within group sum of squares versus the number groups, we see two major elbows. One at groups 3-4 and another one at groups 7-8. This is in agreement with our conclusion from the Rˆ2 plot above, where after looking at various clustering metrics, we concluded that number of groups should be around 8-10. Using both of these metrics, we ultimately decide to go with 8 clustering groups for our cluster analysis.



```
## [1] "Countries in Cluster  1"
##  [1] "Switzerland"          "Iceland"              "Denmark"
##  [4] "Norway"               "Canada"               "Finland"
##  [7] "Netherlands"          "Sweden"               "New Zealand"
## [10] "Australia"            "Austria"              "United States"
## [13] "Luxembourg"           "Ireland"              "Belgium"
## [16] "United Arab Emirates" "United Kingdom"       "Oman"
## [19] "Singapore"            "Germany"              "Qatar"
## [22] "France"               "Uruguay"              "Saudi Arabia"
## [25] "Kuwait"               "Uzbekistan"           "Bahrain"
## [28] "Turkmenistan"
## [1] " "
## [1] "Countries in Cluster  2"
##  [1] "Israel"         "Costa Rica"     "Mexico"         "Brazil"
##  [5] "Venezuela"      "Panama"         "Chile"          "Argentina"
##  [9] "Czech Republic" "Colombia"       "Suriname"       "El Salvador"
## [13] "Guatemala"      "Ecuador"        "Bolivia"        "Paraguay"
## [17] "Nicaragua"      "Peru"           "Jamaica"        "Vietnam"
## [21] "Kyrgyzstan"     "China"
## [1] " "
## [1] "Countries in Cluster  3"
##  [1] "Thailand"            "Malta"               "Trinidad and Tobago"
##  [4] "Malaysia"            "North Cyprus"        "Cyprus"
##  [7] "Mauritius"           "Hong Kong"           "Indonesia"
## [10] "Bhutan"             "Philippines"         "Dominican Republic"
## [13] "Mongolia"           "Sri Lanka"
## [1] " "
## [1] "Countries in Cluster  4"
##  [1] "Spain"       "Taiwan"      "Slovakia"    "Japan"       "South Korea"
##  [6] "Italy"       "Kazakhstan"  "Slovenia"    "Lithuania"   "Belarus"
## [11] "Poland"      "Libya"       "Russia"      "Algeria"     "Estonia"
## [16] "Azerbaijan"  "Jordan"      "Romania"     "Portugal"    "Latvia"
## [21] "Macedonia"   "Albania"     "Lebanon"     "Hungary"     "Ukraine"
## [26] "Bulgaria"
## [1] " "
## [1] "Countries in Cluster  5"
##  [1] "Moldova"                 "Croatia"
##  [3] "Kosovo"                  "Turkey"
##  [5] "Pakistan"                "Montenegro"
##  [7] "Serbia"                  "Morocco"
##  [9] "Bosnia and Herzegovina"  "Greece"
## [11] "Tunisia"                 "Palestinian Territories"
## [13] "Iraq"                    "Armenia"
## [15] "Egypt"
## [1] " "
## [1] "Countries in Cluster  6"
## [1] "Nigeria"          "Zambia"           "Mozambique"       "Lesotho"
## [5] "Swaziland"        "Zimbabwe"         "Liberia"          "Congo (Kinshasa)"
## [9] "Sierra Leone"
## [1] " "
## [1] "Countries in Cluster  7"
## [1] "Somaliland region" "Laos"              "Myanmar"
## [4] "Georgia"           "Rwanda"            "Syria"
## [1] " "
## [1] "Countries in Cluster  8"
##  [1] "Honduras"                 "Tajikistan"
##  [3] "Bangladesh"               "Iran"
##  [5] "South Africa"             "Ghana"
##  [7] "India"                    "Sudan"
##  [9] "Haiti"                    "Nepal"
## [11] "Ethiopia"                 "Mauritania"
## [13] "Kenya"                    "Djibouti"
## [15] "Botswana"                 "Malawi"
## [17] "Cameroon"                 "Yemen"
## [19] "Angola"                   "Mali"
## [21] "Congo (Brazzaville)"      "Comoros"
## [23] "Uganda"                   "Senegal"
## [25] "Gabon"                    "Niger"
## [27] "Cambodia"                 "Tanzania"
## [29] "Madagascar"               "Central African Republic"
## [31] "Chad"                     "Guinea"
## [33] "Ivory Coast"              "Burkina Faso"
## [35] "Afghanistan"              "Benin"
## [37] "Burundi"                  "Togo"
## [1] " "
```

Looking at our cluster groups, we see that a majority of major Western/European countries are grouped together in cluster 1, which is no surprise given that these countries are considered to be highly developed.

We also observe that countries that are generally regarded as undeveloped are grouped together in cluster 8, with a large majority of them from the African continent. Interestingly enough, in cluster 7, we see a group of countries that have all had wars/genocides in the past few decades or recently. This upholds our findings from earlier where in our PCA analysis, we observed that countries, which were recently ravaged by violent conflict, had a higher dependency on democracy and government trust for their happiness. We also observe that some countries that are located in the same region (Ex. East Europe or Middle East) are often clustered together. Overall, we conclude that these cluster grouping of countries largely follows what we would expect to see from a cluster analysis that relies on variables relating to happiness, health, freedom, corruption, and GDP per capita.

Factor Analysis

We start our factor analysis by first looking at the correlation matrix of all the indicator variables in the data used to calculate the happines score of a country. As such, we remove columns refering to the happiness score directly as well as any categorical variables that have to do with the name of a specific country as well as its geographic location.

```
##                               Economy..GDP.per.Capita. Family
## Economy..GDP.per.Capita.                          1.00   0.65
## Family                                            0.65   1.00
## Health..Life.Expectancy.                          0.82   0.53
## Freedom                                           0.37   0.44
## Trust..Government.Corruption.                     0.31   0.21
## Generosity                                       -0.01   0.09
## Dystopia.Residual                                 0.04   0.15
##                               Health..Life.Expectancy. Freedom
## Economy..GDP.per.Capita.                          0.82    0.37
## Family                                            0.53    0.44
## Health..Life.Expectancy.                          1.00    0.36
## Freedom                                           0.36    1.00
## Trust..Government.Corruption.                     0.25    0.49
## Generosity                                        0.11    0.37
## Dystopia.Residual                                 0.02    0.06
##                               Trust..Government.Corruption. Generosity
## Economy..GDP.per.Capita.                               0.31      -0.01
## Family                                                 0.21       0.09
## Health..Life.Expectancy.                               0.25       0.11
## Freedom                                                0.49       0.37
## Trust..Government.Corruption.                          1.00       0.28
## Generosity                                             0.28       1.00
## Dystopia.Residual                                     -0.03      -0.10
##                               Dystopia.Residual
## Economy..GDP.per.Capita.                   0.04
## Family                                     0.15
## Health..Life.Expectancy.                   0.02
## Freedom                                    0.06
## Trust..Government.Corruption.             -0.03
## Generosity                                -0.10
## Dystopia.Residual                          1.00
```

Most of the indicator variables have moderate correlation with each other (less than an absolute value of 60%), but there are some strong exceptions. For example, we see that the amount that family contributes in overall happiness has a relativelt strong correlation of 0.65 with the degree that the economy contributes to happiness. Even stronger correlation is found between the level that health contributes to happiness and the level that is contributed by the economy. This is the highest between any two indicators and it is found to be equal to 0.82. Finally, we can say that we see some degree of moderate correlation between family and health (0.53) as well as freedom and family (0.44).

We continue by performing a Kaiser-Meyer-Olkin (KMO) measure of adequacy test, in order to see if our data are actually appropriate to be used in a factor analysis scheme.

```
## $KMO
## [1] 0.66661
## 
## $MSA
##                                   MSA
## Economy..GDP.per.Capita.      0.61260
## Family                        0.76427
## Health..Life.Expectancy.      0.66260
## Freedom                       0.72968
## Trust..Government.Corruption. 0.68447
## Generosity                    0.52056
## Dystopia.Residual             0.45471
## 
## $Bartlett
## [1] 383.4
## 
## $Communalities
##                               Initial Communalities Final Extraction
## Economy..GDP.per.Capita.                   0.755439          1.08861
## Family                                     0.485684          0.52141
## Health..Life.Expectancy.                   0.683322          0.63214
## Freedom                                    0.428279          0.81384
## Trust..Government.Corruption.              0.296702          0.33251
## Generosity                                 0.224285          0.28501
## Dystopia.Residual                          0.046943          0.18872
## 
## $Factor.Loadings
##                                   [,1]      [,2]     [,3]
## Economy..GDP.per.Capita.      0.940822 -0.432418  0.12838
## Family                        0.679419 -0.098478 -0.22383
## Health..Life.Expectancy.      0.757390 -0.216555  0.10771
## Freedom                       0.676309  0.574811 -0.16136
## Trust..Government.Corruption. 0.440374  0.352875  0.11858
## Generosity                    0.213781  0.462320  0.15989
## Dystopia.Residual             0.065319 -0.080061 -0.42195
## 
## $RMS
## [1] 0.017104
```

Performing a KMO test on our data yields a result of .667, which would put it near the “middling” range for whether factor analysis is appropriate for our data. Even though a value of .667 is lower than what we would like, we can still perform factor analysis on our data as its KMO value doesn’t fall below .50, which is the unacceptable range.

Now, in order to determine the number of latent factors in our model we will be using the results from the principle components analysis that we performed earlier in our report. Specifically, we will be using the same scree plot that we produced earlier in order to determine the number of factors.

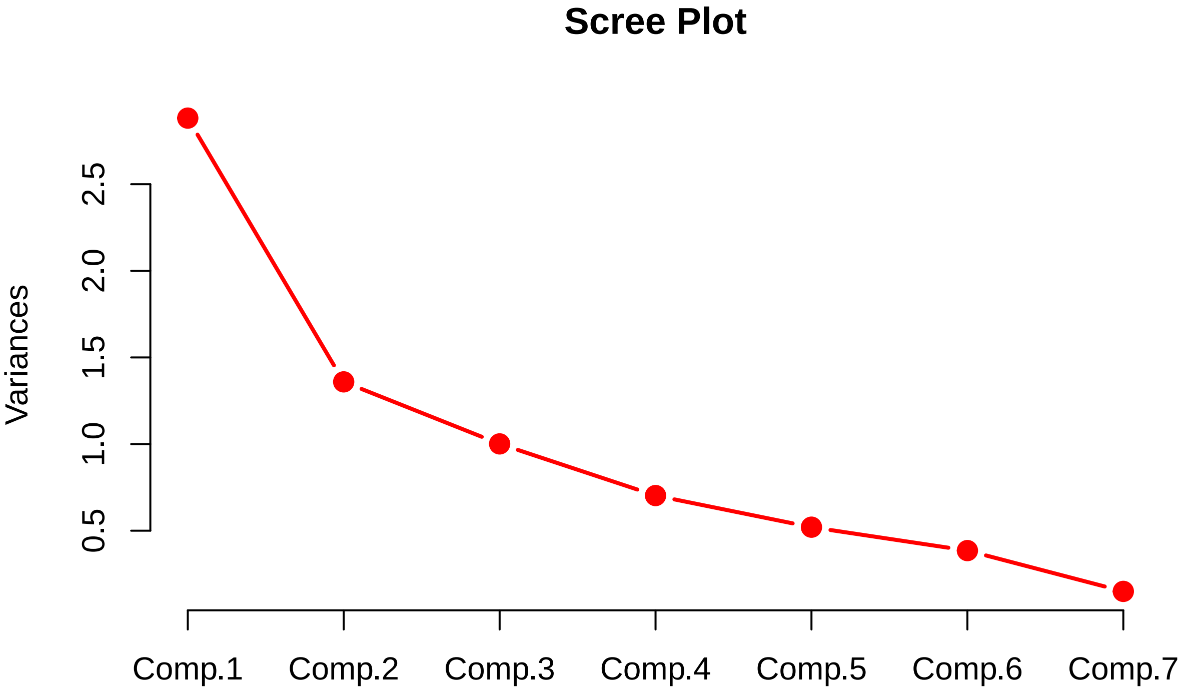
```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"
```

```
## Importance of components:
```

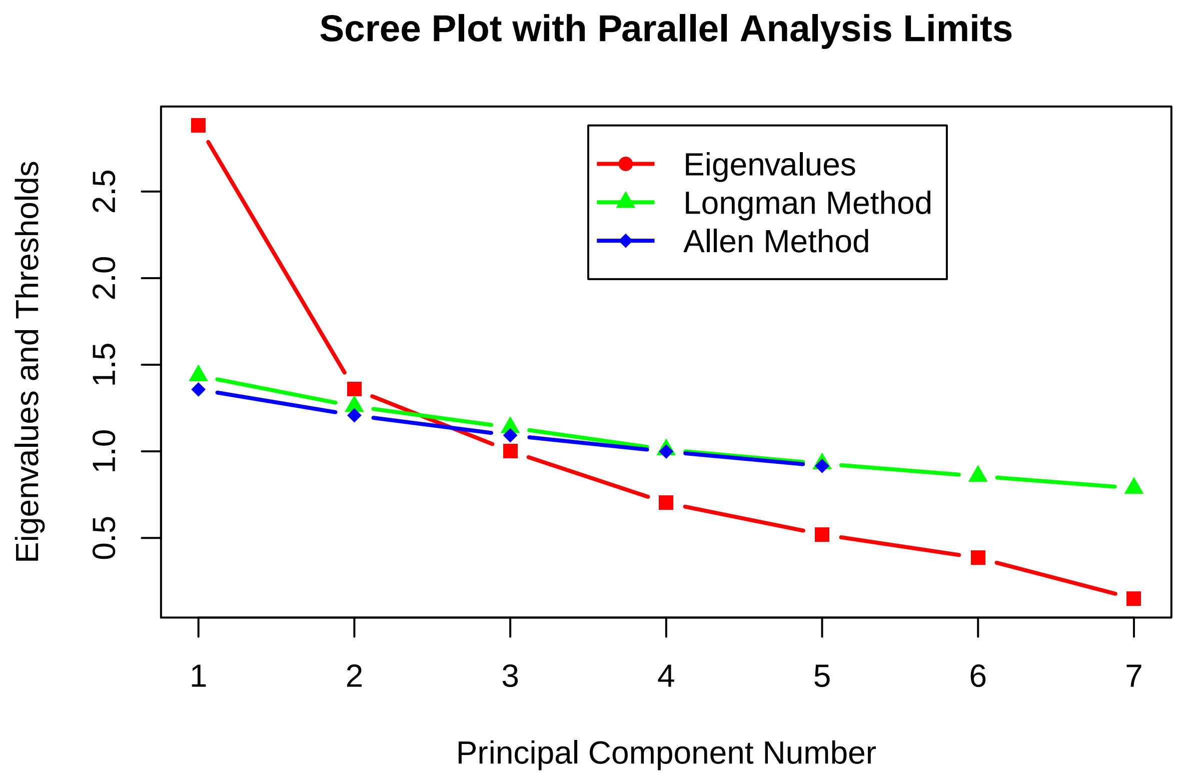
```
##                         Comp.1  Comp.2  Comp.3  Comp.4   Comp.5   Comp.6
## Standard deviation     1.69751 1.16583 1.00055 0.83837 0.721352 0.620695
## Proportion of Variance 0.41165 0.19417 0.14301 0.10041 0.074336 0.055037
## Cumulative Proportion  0.41165 0.60581 0.74883 0.84924 0.923572 0.978610
##                         Comp.7
## Standard deviation     0.38695
## Proportion of Variance 0.02139
## Cumulative Proportion  1.00000
```

```
## Warning in if (loadings) {: the condition has length > 1 and only the first
## element will be used
```

```
##
## Loadings:
##                                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Economy..GDP.per.Capita.         0.50   0.31   0.21   0.04   0.19   0.01   0.76
## Family                           0.45   0.23  -0.07  -0.24  -0.55  -0.58  -0.21
## Health..Life.Expectancy.         0.48   0.24   0.24  -0.13   0.45   0.29  -0.59
## Freedom                          0.42  -0.34  -0.23   0.05  -0.48   0.65   0.04
## Trust..Government.Corruption.    0.33  -0.42  -0.08   0.72   0.21  -0.36  -0.11
## Generosity                       0.17  -0.64  -0.09  -0.63   0.32  -0.20   0.13
## Dystopia.Residual                0.05   0.30  -0.91  -0.02   0.28   0.01   0.02
```



```
##   pcompnum longman   allen
## 1        1 1.43729 1.35743
## 2        2 1.25969 1.20781
## 3        3 1.13796 1.09248
## 4        4 1.00989 0.99883
## 5        5 0.92872 0.91584
## 6        6 0.85688      NA
## 7        7 0.78759      NA
```



Looking at the scree plot, we observe an elbow at 2 factors. Meanwhile on the parallel plot, we see the red eigenvalue line dips below the Longman and Allen method lines between 2 and 3 factors. Ultimately, we decide to use 2 latent factors for our factor analysis.

As we end up having only 2 latent factors in our factor analysis the extraction method we will be using when using orthogonal models will be that of maximum likelihood.

```
##
## Call:
## factanal(x = data[, -c(1:5)], factors = 2, rotation = "varimax")
##
## Uniquenesses:
##      Economy..GDP.per.Capita.                        Family
##                         0.005                         0.531
##      Health..Life.Expectancy.                       Freedom
##                         0.325                         0.157
## Trust..Government.Corruption.                    Generosity
##                         0.703                         0.787
##             Dystopia.Residual
##                         0.997
##
## Loadings:
##                               Factor1 Factor2
## Economy..GDP.per.Capita.       0.994
## Family                         0.626   0.279
## Health..Life.Expectancy.       0.809   0.140
## Freedom                        0.300   0.868
## Trust..Government.Corruption.  0.269   0.474
## Generosity                             0.459
## Dystopia.Residual
```

```
##
##                Factor1 Factor2
## SS loadings      2.200   1.295
## Proportion Var   0.314   0.185
## Cumulative Var   0.314   0.499
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 18.33 on 8 degrees of freedom.
## The p-value is 0.0189
```

```
##                               Economy..GDP.per.Capita. Family
## Economy..GDP.per.Capita.                             0   0.00
## Family                                               0   0.53
## Health..Life.Expectancy.                             0  -0.01
## Freedom                                              0   0.01
## Trust..Government.Corruption.                        0  -0.09
## Generosity                                           0  -0.01
## Dystopia.Residual                                    0   0.11
##                               Health..Life.Expectancy. Freedom
## Economy..GDP.per.Capita.                          0.00    0.00
## Family                                           -0.01    0.01
## Health..Life.Expectancy.                          0.33    0.00
## Freedom                                           0.00    0.16
## Trust..Government.Corruption.                    -0.04    0.00
## Generosity                                        0.08   -0.01
## Dystopia.Residual                                -0.02    0.01
##                               Trust..Government.Corruption. Generosity
## Economy..GDP.per.Capita.                               0.00       0.00
## Family                                                -0.09      -0.01
## Health..Life.Expectancy.                              -0.04       0.08
## Freedom                                                0.00      -0.01
## Trust..Government.Corruption.                          0.70       0.07
## Generosity                                             0.07       0.79
## Dystopia.Residual                                     -0.06      -0.12
##                               Dystopia.Residual
## Economy..GDP.per.Capita.                   0.00
## Family                                     0.11
## Health..Life.Expectancy.                  -0.02
## Freedom                                    0.01
## Trust..Government.Corruption.             -0.06
## Generosity                                -0.12
## Dystopia.Residual                          1.00
```

```
## [1] 0.050934
```

```
## [1] 0.28571
```

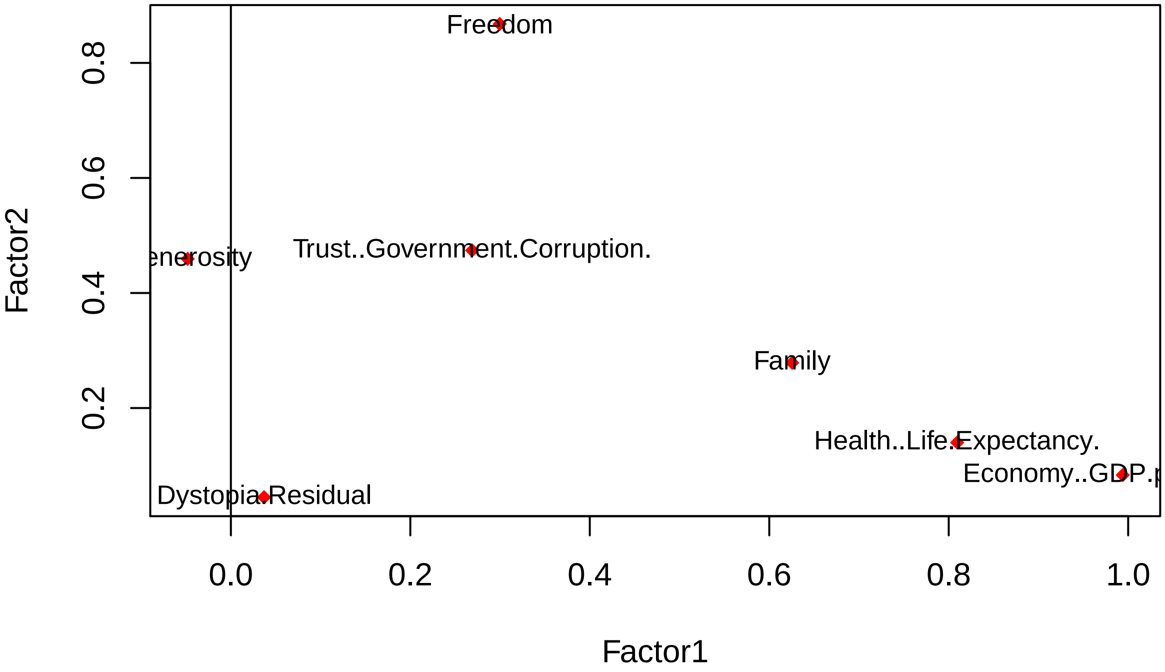
We calculate a Root Mean Square Residual of 0.050934. When we instead look at the proportion of residuals greater than .05, we find a value equal to 0.28571, which is relatively good number.

Next, we will be testing the hypothesis that there are no latent factors in our data.

```
##
## Call:
## factanal(x = data[, -c(1:5)], factors = 2, rotation = "varimax")
##
## Uniquenesses:
```

```
##      Economy..GDP.per.Capita.                            Family
##                         0.005                             0.531
##      Health..Life.Expectancy.                           Freedom
##                         0.325                             0.157
## Trust..Government.Corruption.                        Generosity
##                         0.703                             0.787
##             Dystopia.Residual
##                         0.997
##
## Loadings:
##                               Factor1 Factor2
## Economy..GDP.per.Capita.       0.994
## Family                         0.626   0.279
## Health..Life.Expectancy.       0.809   0.140
## Freedom                        0.300   0.868
## Trust..Government.Corruption.  0.269   0.474
## Generosity                             0.459
## Dystopia.Residual
##
##                Factor1 Factor2
## SS loadings      2.200   1.295
## Proportion Var   0.314   0.185
## Cumulative Var   0.314   0.499
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 18.33 on 8 degrees of freedom.
## The p-value is 0.0189
```

Looking at our p-value of 0.0189, which is less than .05, we can reject the null hypothesis that there are no latent factors in our data. From what we observed in previous steps, we decide to go with 2 latent factors in our factor analysis.



In our loading plot of two factors, we observe that variables such as Economy and Health have particularly high values in factor 1. As both of them are related to a degree with the quality of the instutions in a country, we can say that indicators that are high in value in factor 1, relate to the infrastructure and instutions of a particular country. Meanwhile indicators such as Freedom, Trust and Government Corruption and Generosity have high values in the second factor. Hence, we can say that this factor relates to things such as personal freedoms and quality of interactions among people in a country.

Finally, we can say that the two latent factors that we found relate very closely to the PCA components that we found in the previous part, meaning that the direction towards which the variability of the data increases resembles very closely the latent factorts that describe them.

Conclusions and Discussion

Principal Components Analysis

In our Principal Components Analysis, after using various methods, we decided to keep 2 principal components in our analysis. Using a biplot, our main finding was that countries that are largely dependent on variables like GDP per capita and life expectancy for their happiness were less reliant on variables like generosity and freedom. The same can be said vice versa. Interestingly in our PC score plot, we found that countries that have experienced recent violent events like Syria, Rwanda, and Myanmar were negative outliers towards the 2nd component. This means that these countries are heavily reliant on freedom and generosity for their happiness levels.

Cluster Analysis

Using a euclidean distance and ward.D agglomeration method, we grouped countries into 8 clusters. We largely saw what we expected to see from a cluster analysis where countries were grouped by variables related to happiness, health, freedom, corruption, and GDP per capita. For example, we had a cluster of Western countries that many would consider as highly developed countries. Additionally, we also had a cluster of countries that would be considered as undeveloped and low-income, largely in the African continent. Suprisingly, we also observed a cluster that consisted of countries that have experienced war and genocide in the past decades, supporting our findings earlier in our PCA analysis.

Factor Analysis

Factor analysis gave us some interesting insights that would not be possible with the other two multivariate methods. Through PCA we showed that there are indeed two factors towards which the variability in our data increases and along these axes our variables differ. Here we built on top of that and were able to get an insight on what the latent variables are. These were able to describe the happiness score of each country in a lower dimensional space. Public infrastructure and personal freedoms are two things that someone could easily associate with happiness of people in a country. Hence, it would be interesting to see future happiness reports expand upon the variables that they already include or group them into fewer categories as we showed.

References

The introductory part of this project as well as the descriptions for the variables were adapted from this page, which features and in-depth overview of the contents of the dataset: https://www.kaggle.com/unsdsn/world-happiness/data#